DOCUMENT RESUME

ED 388 714 TM 024 161

AUTHOR Donoghue, John R.; Jenkins, Frank
TITLE A Monte Carlo Study of the Effects of Model Misspecification on HLM Estimates.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-92-69
PUB DATE Nov 92
NOTE 44p.; Version of a paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Correlation; *Estimation (Mathematics); Models; Monte Carlo Methods; Predictor Variables; Sample Size
IDENTIFIERS *Hierarchical Linear Modeling; *Specification Error

ABSTRACT

Monte Carlo methods were used to investigate the effect of misspecification of the second level in a two-level hierarchical linear model (HLM). Sample composition, heterogeneity of the group size, level of intraclass correlation, and correlation between second-level predictors were manipulated. Each of 20 generated data sets was analyzed nine times with the HLM program, corresponding to the correct model and eight types of misspecification. The error variance was estimated accurately for all model specifications. For the correct second-level model, regression parameters and the covariance matrix of the second level errors were estimated accurately, although estimates were slightly biased for some levels of the sample composition manipulation. Misspecifications that failed to include a predictor had a larger effect than those which erroneously included an effect. Reactivity was assessed by examining parameters of equations which were correctly specified, although another equation in the model was misspecified. HLM showed little reactivity. Only one of the misspecified models yielded estimates that were significantly farther, on average, from the generating parameters than were those of the correctly specified model. An appendix contains three tables describing the correctly specified model. (Contains 12 text tables and 21 references.) (Author/SLD)

RR-92-69

RESEARCH REPORT

A MONTE CARLO STUDY OF THE EFFECTS OF MODEL MISSPECIFICATION ON HLM ESTIMATES

John R. Donoghue
Frank Jenkins

Educational Testing Service
Princeton, New Jersey
November 1992

A Monte Carlo Study of

The Effects of Model Misspecification on HLM Estimates

John R. Donoghue and Frank Jenkins

Educational Testing Service

An earlier version of this paper was presented at the annual meeting of the American Educational Research Association in San Francisco, CA April 1992. The authors wish to thank Jim Carlson and John Mazzeo for their numerous helpful comments, Donna Lembeck for her assistance in preparing tables, Richard Congdon for his assistance in implementing a Sun version of the HLM program, and Mike Wagner for his help with the programs and the Sun operating system.

Abstract

Monte Carlo methods were used to investigate the effect of misspecification of the second-level in a two-level hierarchical linear model. Sample composition, heterogeneity of the group size, level of intraclass correlation, and correlation between second-level predictors were manipulated. Twenty data sets were generated for each condition. Each data set was analyzed nine times with the HLM program, corresponding to the correct model and eight types of misspecification.

The error variance σ^2 was estimated accurately for all model specifications. For the correct second-level model, regression parameters γ and the covariance matrix of the second-level errors τ were estimated accurately, although estimates were slightly biased for some levels of the sample composition manipulation. Misspecifications that failed to include a predictor had a larger effect than those which erroneously included an effect.

Reactivity was assessed by examining parameters of equations which were correctly specified, although another equation in the model was misspecified. HLM showed little reactivity. Only one of the misspecified models yielded γ estimates that were significantly farther, on average, from the generating parameters than were those of the correctly specified model. Similar results were found for the root mean square residual between the generating and estimated τ.

Hierarchical Data

Much educational research takes place in hierarchically ordered settings. Naturally, pupils are found grouped into classes, classes occur within schools, schools within districts, districts within counties and states, and states within nations. Attempts to study or to modify educational practice may occur at any of these levels, and the targets of these efforts often occur at other levels. Traditionally, the hierarchical nature of educational data has been difficult to accommodate. One of the chief limitations was that standard statistical techniques do not adequately account for the dependencies in hierarchical data. Faced with this dilemma, researchers chose one of two options. Either they ignored the dependencies within the data and analyzed individual data points, or they aggregated data to higher levels of the hierarchy and analyzed means. The first strategy leads to underestimation of variance estimates (see Raudenbush, 1988 and the references cited therein), and can consequently yield grossly distorted significance tests. The second strategy leads to aggregation bias, in that relationships (e.g., correlations) of means may be markedly different from the same relationships based on the unaggregated variables (again, see Raudenbush, 1988 and the references cited therein). Furthermore, aggregation changes the meaning of variables; the average class achievement score does not represent the same educational process as individual achievement.

In the 1980's several papers proposed methodologies for hierarchical data. Specifically, Aitkin and Longford (1986), DeLeeuw and Kreft (1986), Goldstein (1986), Mason, Wong, and Entwisle (1984), Raudenbush (1988), and Raudenbush and Bryk (1986) all proposed methods of estimating effects within hierarchical contexts. These approaches enabled researchers to model educational processes at several levels of aggregation thereby addressing issues of aggregation bias, efficient estimation of effects, and individual by setting interactions (Raudenbush, 1988).

This paper focuses on the two-level hierarchical linear model utilized by the HLM program (Bryk, Raudenbush, Seltzer, & Congdon, 1989). For consistency, we adopt their notation. The model begins with the first-level--a within-group multiple regression model to predict the outcome variable:

$$y_{ij} = X_{ij}\beta_j + R_{ij} \tag{1}$$

in which the outcome for person i in group j, y_{ij}, is a function of an individual's characteristics. R_{ij} is a normally distributed (iid over persons and groups) error term. β_j is a vector of regression coefficients that are considered to be random effects. For example, within a school, students' math achievement could be predicted by student SES and ethnicity (Lee, 1986). The second-level, between-group model is given by:

$$\beta_j = W_j\gamma + U_j \tag{2}$$

in which the parameters from the first-level model are functions of group-level characteristics. For example, the SES to achievement relationship could be a function of school sector (i.e., public or private, see Lee, 1986). The U_j terms are assumed to be independent of R_{ij}, but the U_j for different β's are not assumed to be independent of each other. The U_j terms are jointly multivariate normal with mean vector 0 and covariance matrix τ. The vector of γ's are usually considered to be fixed effects. It should be noted that the data matrices X_{ij} and W_j are specially structured to accommodate the hierarchical nesting of the data (see Strenio, Weisberg, & Bryk, 1983).

Estimation of hierarchical models is complex. The HLM program (Bryk et al., 1989) uses an empirical Bayes approach for estimating structural parameters. Variance estimates are obtained via the expectation-maximization (EM) method (see Dempster, Laird, & Rubin, 1977).

When these approaches are combined, they yield full information maximum likelihood estimates of fixed and random effects (see Dempster, Rubin, & Tsutakawa, 1981). Technical details may be found in Raudenbush and Bryk (1985), Bryk et al. (1989), and Bryk and Raudenbush (1992).

As hierarchical linear models gain wider use in applied research, it is important to examine their properties in order to properly evaluate the empirical applications of the models. Currently, hierarchical linear models are being used to inform educational practice pertinent to a wide variety of issues such ar school effectiveness (Aitken & Longford, 1986), school policy (Raudenbush & Bryk, 1986), and individuals' growth in learning (Bryk & Raudenbush, 1987). Such studies usually employ pre-existing, non-experimental data. Substantive inferneces in such contexts are problematical. With hierarchical analyses the situation is further exacerbated by the fact that imperfectly specified linear models are posited for at least two levels of aggregation and estimated with full information methods. Given this combination of complexity and full information estimation, many researchers have worried that such models may be very reactive, i.e., that the addition or deletion of one or two predictors will lead to very different estimates of all other effects. If this is the case, hierarchical analyses would be poor tools for informing educational policy and practice. The present study examines how prone hierarchical linear models are to reactivity;Under what conditions are estimates robust to misspecification?

Model Misspecification

Some of the effects of model misspecification may be guessed from a general knowledge of multiple regression. For the purposes of discussion, assume that this equation:

$$\beta_q = \gamma_{0q} + \gamma_{1q} W_{1q} + \gamma_{2q} W_{2q} + \ . \ . \ . \ + \gamma_{pq} W_{pq} + U_q \qquad (3)$$

represents a regular, single-level multiple regression. Consider the effect of incorrectly omitting a predictor W_{mq} from the prediction of β_q (i.e., setting $\gamma_{mq} = 0$). We know that the effect of this misspecification depends on the correlation of W_{mq} with other predictors W_{kq} . If W_{mq} is uncorrelated with other predictors, the effect will be to reduce R^2 by an amount directly proportional to the correlation between β_q and W_{mq} , and the other parameters, γ_{kq} , will be unaffected. On the other hand, if W_{mq} is correlated with W_{kq} , R^2 will not decrease as much, but γ_{kq} will increase or decrease, depending on the sign of the correlation.

In the HLM context, this intuitive analyses is complicated by the nature of the model. β_q is not an observed variable, it is a regression coefficient and its value depends on the predictor variables in the first-level equation. In addition, the errors in the second-level model are not independent but correlated; misspecification in one equation may affect estimates in other equations. Because HLM employs a full information technique of estimation, a single change in specification can, in principle, have effects on every other parameter. Bryk and Raudenbush (1992) describe several features which may affect the estimates of hierarchical linear model parameters. Specifically, they demonstrate that the parameters throughout the model may be affected by a single misspecification at the second-level of the model.

To investigate the reactivity of HLM, we conducted a simulation study. The levels of conditions which were crossed were chosen to represent a range of conditions that are typical for school based research in which students are nested within classes or schools (e.g., the reanalysis of the High School and Beyond data , Lee, 1986). We hoped to determine under what conditions the analysis is very reactive to the sort of model misspecification that commonly occurs in naturalistic studies.

Method

Data were generated according to a two level HLM model. A single true model was used to generate the data. Features of the second-level model were varied in several ways; characteristics of the first-level of the model were not varied. The basic model contained two predictors at the first-level and three predictor variables at the second-level. In order to increase the relevance of the study, parameters of the basic model were similar to those obtained by the authors in an empirical HLM analysis of the High School and Beyond (Coleman, Hoffer & Kilgore, 1982) data.

Design

The chief variable of interest was the specification of the second-level HLM model for analysis. In generating the data, four other "nuisance variables" were manipulated, due to their high salience. The independent variables were:

1) Sample composition (5 levels). This factor was jointly determined by the number of subjects within a group and the number of groups. The number of subjects and number of groups were manipulated to define 5 conditions such that the expected total number of subjects was 1500 for each data set: (a) 10 groups with an average of 150 subjects in each group; (b) 25 groups with an average of 60 subjects in each group; (c) 60 groups with an average of 25 subjects in each group; (d) 150 groups with an average of 10 subjects in each group; and (e) 300 groups with an average of 5 subjects in each group.
2) The heterogeneity of the number of subjects within a group (2 levels)-- low and high heterogeneity.
3) The correlation between second-level predictors, W (2 levels)--0.30 and 0.60.

4) The intraclass correlation of the observations (2 levels)--0.30 and 0.60.

5) Specification of the model for HLM analysis (9 levels)--the correctly specified model and eight types of misspecification, including errors of omission and erroneous inclusion of effects. These models are schematically presented in Table 1.

The first four factors were fully crossed to yield 40 conditions. Twenty data sets were generated for each condition, according to the procedure given below. Each data set was then analyzed according to the 9 model specifications indicated in Table 1, yielding a total of 7200 HLM analyses.

Insert Table 1 about here

Data Generation

To simplify the presentation, we have modified the usual HLM two-level model notation. Particularly, in discussing the second-level equations, some of the usual the notation associated with the first-level model is suppressed.

The first step in generating the data was to generate the group (second) level data. For each group, the vector of group level parameters β_j was generated according to equation (2). To simplify the study, the second-level effects **W** were sampled from an appropriate distribution, the parameters of which were chosen to be similar to the 1988 HSB data:

$$\mathbf{W}_j \sim N(0, \Xi)$$

$$\Xi = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

$$\mathbf{U} \sim N(0, \tau) \qquad (4)$$

$$\tau = \begin{bmatrix} 6.6 & -1.5 & 1.2 \\ -1.5 & 9.4 & -1.6 \\ 1.2 & -1.6 & 2.5 \end{bmatrix}$$

where ρ was determined by the value of the third design factor, the correlation between the second-level predictors. The second-level effects were specified:

$$\begin{aligned} \beta_0 &= 10.07 + 15.27\, W_1 + 5.51\, W_2 + U_1 \\ \beta_1 &= 17.52 \qquad\qquad - 1.56\, W_2 + 5.59\, W_3 + U_2 \\ \beta_2 &= 3.69 + U_3 \end{aligned} \qquad (5)$$

The next step was to generate the subjects within groups (first-level data). For a given sample composition, the corresponding average number of subjects per group was used. For each group, the n_j was sampled as the nearest integer from a modified beta distribution, B(r,s):

$$n \sim \bar{n} * (B(r,s) + 0.5) \qquad (6)$$

with r equal to s. Thus, group size ranged from 0.5 to 1.5 times the average number of subjects per group. The parameters r and s were chosen according to the heterogeneity condition, r=s=1 (uniform distribution) for the high heterogeneity condition and r=s=4 (peaked, r=s=1 (uniform distribution) for the high heterogeneity condition and r=s=4 (peaked, symmetric distribution) for the low heterogeneity condition.

Finally, the outcome variable was generated for each subject in each group according to equation (1). As was the case for the second-level data, the fixed effects X were sampled from a normal distribution, the parameters of which were chosen to be similar to the 1988 HSB data:

$$X \sim N(\mathbf{0}, \mathbf{\Delta})$$

$$\mathbf{\Delta} = \begin{bmatrix} 1.0 & 0.2 \\ 0.2 & 1.0 \end{bmatrix} , \qquad (7)$$

and the individual errors R_{ij} were iid, sampled from a $N(0,\sigma^2)$ distribution. The first-level error variance was determined from the intraclass correlation ρ_I and the second-level variance of β_0, τ_{11}:

$$\sigma^2 = \frac{\tau_{11} - \tau_{11}\rho_I}{\rho_I} . \qquad (8)$$

Data were generated by a FORTRAN 77 program written by one of the authors. Uniform pseudo-random numbers were generated using the algorithm by Wichman and Hill (1982). Univariate normal deviates were obtained by transforming uniform deviates with the inverse normal distribution function transformation of Beasley and Springer (1977). Multivariate normal variables were obtained by generating independent normal deviates, and then multiplying by the Cholesky decomposition of the desired population covariance matrix. Beta deviates (for sampling n_j) were obtained by transforming a pair of gamma deviates, using the algorithm given in Press, Flannery, Teukolsky, and Vetterling (1989).

For group j, n_j first-level observations were generated according to equation (1). The sufficient statistics (sums and sums of squares and cross-products) were then calculated for that group.[1] Finally, the sufficient statistics for each group were written to a binary file for input into the HLM program.

Dependent Variables

We wanted to assess the recovery of two different types of parameters, the second-level regression coefficients γ, and the variance components σ^2 and τ. In the first case, we were concerned with the discrepancy between the estimated and true values of γ. In predicting β_q, the discrepancy was measured by the euclidean distance between the estimated vector and the true vector.

$$d_q = \sqrt{\sum_{m=0}^{p} (\hat{\gamma}_{mq} - \gamma_{mq})^2} \quad . \tag{9}$$

For misspecified models, non-zero elements of γ which were constrained to be zero (i.e., errors of omission) were excluded from the distance.

The second focus of interest was the accuracy of estimates of the variance components, the covariance matrix of second-level errors τ, and the residual variance σ^2. In the context of predicting a single effect β_q, we can differentiate between τ_{qq}, the error variance associated with that effect, and errors associated with other effects, τ_{tu}, t,u $\neq$ q. For comparing submatrices of

[1] While generating the sufficient statistics directly might have been more efficient, it was felt that the method used here was safer. Generating sufficient statistics directly relies heavily on the distributional properties of the generating algorithm. Generating individual level data and then computing means and covariances from that data allows the central limit theorem to function. This acts to strengthen the distributional properties of the generating algorithms.

the τ matrix, the root mean square residual was used:

$$RMR_q = \sqrt{\frac{\sum_{t \neq q} \sum_{u \leq t, u \neq q} (\hat{\tau}_{tu} - \tau_{tu})^2}{p(p+1)/2}} \quad . \tag{10}$$

A simple bias measure was used for a single parameter θ (σ^2 or τ_{qq}):

$$b_\theta = \hat{\theta} - \theta \quad . \tag{11}$$

Analyses

To summarize the results, separate factorial analyses of variance were conducted with each outcome measure serving as the dependent variable. The independent variables were the design factors used to generate the data. Due to the nature of the dependent variables, there is every reason to expect that the assumptions underlying the ANOVA significance test will be violated. A practical criterion of effects accounting for three or more percent of the variance was adopted in favor of traditional significance testing. The statistic computed to assess practical significance was η^2 (e.g., Tabachnick & Fidell, 1983),

$$\eta^2 = \frac{SS_{effect}}{SS_{Tot}} \quad . \tag{12}$$

Results

The results of the study are presented in four sections. First, we characterize some general aspects of the behavior of the HLM program. Specifically, we examine the occurrence of convergence, problems encountered during iteration, and the number of iterations required to yield convergence. Second, we consider the estimation of the first-level error variance, σ^2. Because this parameter seemed particularly insensitive to manipulations, this analysis is broader than those of γ or τ. Next, we consider the ability of the HLM program to recover the generating model for γ or τ. Because any discussion of misspecification is relative to behavior of the correctly specified model, we examine various aspects of this model in some depth. Finally, we consider the effects of several types of misspecification at the second-level of the model.

Behavior of the HLM Program

A) Convergence

Using the default HLM criterion for convergence, we were able to obtain a converged HLM solution for every run in the study. However, several runs exhibited some problems during iteration. These problems fell into two broad categories: problems with the automatic start values, and non-positive definite estimates of the τ matrix during the course of iterations. Accepting the "automatic fix-up," when prompted by the program, was uniformly successful in solving the problem of start values.

Problems with non-positive definite estimates of τ proved somewhat more difficult to handle. Our procedure was to attempt the HLM automatic fix three times. In the majority of cases, this procedure proved successful in yielding a converged solution. In a minority of cases, three attempts using the automatic fix were not successful. In these cases, we next selected the option of setting τ to be diagonal. Our batch program allowed this fix to be tried three times as well. However, it was never necessary to attempt it more than twice in order to yield a converged solution.

B) Number of Iterations

Bryk and Raudenbush (1992) suggest that the number of iterations may be used as a rough guide to the adequacy of the model. In particular, they suggest that poorly defined models may require very large numbers (i.e., > 100) of iterations to converge, while good models typically converge quickly. The vast majority of runs converged in under 30 iterations, and over 99% of the runs had converged by the 60th iteration.

Longer runs were associated with small sample sizes at one of the levels of the model. Fifteen runs required over 100 iterations to converge. All of these long runs occurred for the 300 groups, 5 per group condition. An additional 80 runs from this condition required between 51 and 100 iteration to converge. The second most common condition with longer runs consisted of data sets containing 10 groups with an average of 150 per group; 6 runs required 51 to 100 iterations to converge. The remaining conditions never required more than 50 iterations to converge. Long runs (> 100 iterations) were not strongly associated with any particular model specification; each model specification yielded at least one long run. It should be noted that these results may be influenced by the strategies described above for dealing with problems encountered during estimation.

In an attempt to further characterize long runs, we examined the value of the likelihood function at the start of iterations and at convergence. If the difference between these two values is small, a large number of iterations indicates that the likelihood function is relatively flat, a condition associated with a poorly defined model. On the other hand, if the difference is large, the large number of iterations indicates that the starting values functioned relatively poorly, and that HLM required a large number of iterations because estimation started a long way from the final solution. Table 2 tabulates the average value of the likelihood functions at the first iteration and at convergence for long and short runs. These results support the hypothesis of poor start values; the value of the likelihood is much smaller at the first iteration of long runs, but closer to that of short runs at convergence. However, on average, long runs yield somewhat lower likelihood values than do short runs. The average number of fix-ups for the starting values is also higher for long runs than for short runs. While part of this effect may be due to the strategy we adopted for restarts, there does appear to be some evidence that large numbers of iterations are associated with poor start values.

Insert Table 2 about here

Overall, however, the HLM program functioned well. The vast majority of runs converged rapidly and without problems. Where problems with starting values occurred, automatic fix-ups were successful in yielding convergence. Problems during iteration and long runs were associated with small sample sizes at one of the levels of the model.

Estimation of σ^2

Preliminary analyses indicated that σ^2 was estimated accurately across a wide range of model specifications and conditions. The average value of σ^2 was 9.90, and the overall bias in estimating σ^2 was 0.007. To determine whether any factors in the study affected the estimate, we performed a five-way factorial ANOVA, with the design factors as independent variables and b_σ, the bias in the estimate of σ^2, as the dependent variable. The estimation of σ^2 was not substantially affected by any of the design factors; η^2 for the largest effect is less than .01 and the entire model only accounts for five percent of the variance in σ^2. Further evidence of the stability of σ^2 is given by its small mean squared error. Ignoring the presence of the design factors, the mean squared error of σ^2 is 0.259, compared to its overall mean of 9.90.

Based on these results, σ^2 may best be viewed as a characteristic of the first-level of the HLM model. Because our study was designed to examine the effect of various aspects of the second-level of the model, while holding the first-level constant, it is not surprising that we found no effects for σ^2.

Recovery of the Correctly Specified Model

A logical prerequisite to discussion of misspecification is a good understanding of how HLM functions when the model is correctly specified. Discussions of parameter estimate bias, or mean squared error, under misspecification are only meaningful in comparison to the behavior of the same quantities for the correct model.

Overall, the recovery of the parameters of the correct model was excellent. Table 3 indicates that the point estimates of γ were virtually unbiased for the correct model. Estimates of the second-level error variance τ also functioned very well. The small amount of bias that was present for specific elements does not appear to follow any pattern.

Insert Table 3 about here

Estimation of γ and τ

To investigate the effect of the design manipulations, separate factorial ANOVAs were performed for the dependent variable d (the euclidean distance between the observed and generating γ matrices) and RMR (the root mean square residual between the observed and generating τ matrices).[2] The results indicated that the sample composition had a large effect on

[2] The two measures d and RMR are both based on sums of squared differences. RMR is divided by the number of unique elements in the sum, while d is not. Different measures are used in an attempt to reinforce which parameter matrix, γ or τ, is being discussed.

both dependent variables. None of the other factors had a substantial ($\eta^2 \geq .03$) effect on either dependent variable.

Table 4 gives the mean d and RMR for each sample composition (identified by the number of groups), as well as the mean for each of the γ and τ coefficients. There are no clear patterns in the results for individual parameters. On the other hand, both d and RMR declined monotonically as the number of groups increased, indicating that the quality of the γ and τ estimates improved as the number of groups increased. The standard deviations of both d and RMR also decreased as the number of groups increased. This indicates that the variability in the quality of the estimates decreased; there were fewer very bad solutions as the number of groups increased. (Table A1 in the Appendix gives results of the full ANOVAs, and Tables A2 and A3 give means for all of the main effects in the design. In general, there was little effect for any of these manipulations.)

Insert Table 4 about here

Estimates of se(γ_{mq}) for the Correct Model

For each of the twenty replications within a cell of the design, the empirical standard deviation of the estimates was computed (σ), as was the average of the standard error estimates (se). The ratio se/σ was formed to evaluate the average accuracy of the estimated standard error of γ within each cell. It should be borne in mind that the variance in estimating σ^2 and τ is not reflected in the estimate of the variance of the γ's. Mean ratios (averaged over cells) for the standard error of each of the γ parameters in the correctly specified model are presented in Table 5. There is a tendency for standard errors of γ's associated with slopes (as opposed to base coefficients) to be underestimated by about five to ten percent; standard errors of base terms were overestimated by two to seven percent. Overall, the standard error estimates appear to function well. However, there was considerable variability in the accuracy of the estimated standard errors, as is indicated by the relatively large range and standard deviation of the ratios. Further study of these estimates would appear to be warranted.

Insert Table 5 about here

Effects of Model Misspecification

In examining the effects of model misspecification, we made a distinction between the incorrect second-level equation (the studied equation) and the other two second-level equations. For example, in Model D in Table 1, γ_{11} is erroneously included. Thus, the equation predicting β_1 is the studied equation. For misspecified models, two distance measures were computed. The first, dm_s (the distance of the vector of γ's from the parameter values for the misspecified model) was computed over γ's in the studied equation. For example, in Model D, dm_s included γ_{01}, γ_{11}, γ_{21}, and γ_{31}. Recall that dm_s did not include elements of γ that had been erroneously constrained to 0; the measure only includes γ's that, in principle, could have been estimated correctly. The second distance measure, dm_o, was computed over the other two equations. In Model D, for example, dm_o included γ_{00}, γ_{10}, γ_{20}, and γ_{02}. Similarly, two measures of the τ matrix were computed. The bias measure bm_τ (bias in τ for the misspecified model) was computed for the studied equation (τ_{22} for Model D, for example). $RMRm_o$ was the root mean residual computed over the unique elements of τ that did not involve the studied equation (for example, τ_{11}, τ_{13}, and τ_{33} in Model D).

As a comparison for a given misspecified model, the distance measure of the corresponding correctly specified equation, dc_s, was computed. Similarly, the RMR of the elements of τ not in the studied equation, $RMRc_o$, was computed for the correctly specified model. To evaluate the effect of misspecification, the difference was computed between the misspecified model and the correctly specified model, e.g. difference = $dm_o - dc_o$. Positive differences indicate that, on average, the misspecified model was farther from the generating parameters than the correctly specified model was from the generating parameters. The difference compares the relative closeness of the two estimates to the true parameters; it does not provide any direct information about how similar the two solutions were to each other.

Estimates of γ for the Studied Equation

For each of the nine analyses of a data set, the euclidean distance between the estimated and generating values of γ for the studied equation, was computed according to Equation (9). Table 6 summarizes the results of the distance measures for the studied equation. Positive values indicate that the estimates from the misspecified model were farther from the generating values than were the estimates from the correctly specified model. Misspecification in the second-level model leads to estimates of γ that are, on average, farther from the generating values that are γ estimates derived for the correct specification. Not all misspecifications are equal, however. Specifications A and C, which both include errors of omission, are much farther from the generating values than were models which contained only errors of commission. This is to be expected; when an effect is omitted, regression coefficients of other, correlated predictors increase. However, Model E, which also contained an error of omission, did not show this effect as strongly as did Models A and C.

Insert Table 6 about here

To examine the effects of the design factors on the distance measures for the studied equation, separate factorial ANOVAs were performed for each model specification. The design factors were the independent variables. The difference in the distances, $dm_s - dc_s$, served as the dependent variable. The criterion $\eta^2 \geq .03$ was used to identify salient effects. The most commonly identified effects were the sample composition and the correlation between W's. Means for these main effects are given in Table 7. In general, as the number of groups increases, the effect of misspecification decreases. For a few cases, the 300 group condition yields slightly higher means than does the 150 group condition, but the difference is never large. As would be expected from standard multiple regression, the effect of misspecification is larger for a correlation among W's of 0.6 than for a correlation of 0.3.

Insert Table 7 about here

Estimates of γ for the Other Equations

For each of the nine analyses of a data set, the euclidean distance between the estimated and generating values of γ for the other equations, was computed according to Equation (9). Table 8 summarizes the results of the distance measures for the other equations. Thus, these results address the reactivity of second-level regression parameters in the HLM model. If the model is very reactive, we would expect the distance measures for misspecified models to have much larger means than those of the correctly specified model. The results indicate that, for the misspecifications examined, HLM is not very reactive; only Model E exhibits a significantly larger distance measure than the correct model. Model A exhibits the pattern to a small extent, but the difference is not statistically significant. On the other hand, all of the other models exhibit the opposite pattern; on average, the other elements of γ are somewhat closer to the generating parameters for misspecified models than for correctly specified models.

Insert Table 8 about here

To examine the effects of the design factors on distance measures for the other equations (that do not contain a misspecified γ), separate ANOVAs were performed for each model. The design factors were the independent variables, and the difference dm_o - dc_o served as the dependent variable. For the majority of models, none of the effects yielded $\eta^2 \geq .03$. In the few cases that there was a salient effect, it was always for the sample composition factor. Means for this main effect are given in Table 9. In general, the 10 groups condition is less like the correctly specified model than are any of the other conditions.

Insert Table 9 about here

Estimates of τ for the Studied Equation

Table 10 summarizes the results of the bias measure b_τ for the studied equation, i.e., the bias in τ_{qq} computed from Equation (11). As was the case for γ, the results for the correctly specified model are also given. In general, the results are as expected. Errors of omission in the specification of the second-level model (Models A, C and E) lead to estimates of τ_{qq} that are larger than the generating values. Omissions on the intercept term β_0 yield much larger differences (errors of 24-25 times τ_{11}) than do errors on β_1, which are only about 20 percent of τ_{22}. On the other hand, misspecified models that contain only errors of commission result in very little bias in τ_{qq}, and yield estimates which are very similar to those derived for the correct specification.

Insert Table 10 about here

Separate factorial ANOVAs were performed for each model specification. The design factors were the independent variables, and the difference bm_r - bc_r served as the dependent variable. The only effect that was identified as having a salient effect was the correlation between W's. Means for this main effect are given in Table 11. As would be expected from standard multiple regression theory, the effect of misspecification was smaller when the correlation was 0.6 than when it was 0.3. When predictors are highly correlated, the effect of omitting a predictor may be mitigated by increasing the weights of other, correlated predictors. As the correlation decreases, so does this effect, and errors of omission have correspondingly larger effects in increasing the error variance.

Insert Table 11 about here

Estimates of τ for the Other Equations

Table 12 summarizes the results of the RMR measures for the τ's derived from the other equations, i.e., second-level equations with no misspecified γ's. Thus, these results address the reactivity of second-level τ parameters in the HLM model. If the model is very reactive, we would expect the RMR measures for misspecified models to have larger means than those of the correctly specified model. The results indicate that, for the misspecifications examined, HLM is not very reactive. None of the models yield an RMR which is significantly different from that of the correctly specified model.

Insert Table 12 about here

Separate factorial ANOVAs were performed for each model specification. The design factors were the independent variables, and the difference $RMRm_o$ - $RMRc_o$ served as the dependent variable. None of the effects yielded $\eta^2 \geq .03$ for any of the models.

Discussion

There are several limitations to these results. Probably the strongest limitation was the use of a single first-level model. This model was chosen to be very similar to an empirical HLM analysis, and the parameters were chosen to closely resemble a substantitively interesting examination of the HSB data. Nonetheless, it may be that aspects of the first-level model, such as the correlation between the first-level predictors, can have a strong influence on our results. The extent to which our findings depend upon the characteristics of that model are unknown at present; we are currently extending our work along these lines. The use of a single second-level model is a less serious limitation, as there is little reason to expect our findings to be limited to the specific model used; aspects of this model were varied, and the lack of strong effects for those manipulations gives us some confidence that our results are not overly limited by this aspect.

The sample composition had salient ($\eta^2 \geq .03$) effects on several of the outcome measures. The manner in which we manipulated this variable argues against the seemingly plausible assertion that adding subjects to each group will compensate of a lack of groups. While this line of reasoning holds for some aspects of hierarchical models (e.g., estimating the intraclass correlation, Collins, Donoghue & McGuigan, 1989), we found no evidence that it applies to HLM quantities specifically associated with the second-level model such as γ or τ.

Bryk and Raudenbush (1992) have discussed this characteristic of HLM; our results support them.

Two of the manipulations, the value of the conditional intraclass correlation and the heterogeneity of group sizes, had no large effects for any of the dependent variables we examined. These findings differ from those of Bassiri (1988), who found effects for both of these manipulations. It is not known whether this discrepancy is due to differences in methodology, or is a function of the specific model specifications used in the two studies. Further research will be required to determine the answer to this question.

One limitation of our results is a failure to identify conditions under which HLM clearly broke down. Our design was selected to span a range of commonly occurring situations in educational research. Little was known about the behavior, especially reactivity, of HLM across this range. It was reassuring to find that the program functioned well across this range. Nonetheless, knowledge about where the model clearly breaks down would also be valuable.

In general, our results are very encouraging. HLM estimates were unbiased for the correctly specified model, and showed expected patterns of results for parameters associated with the equation which was directly affected by the misspecification. Moreover, we saw little evidence of reactivity, despite HLM's use of full information techniques of estimation. The extent to which our findings may be extended to other HLM designs is not clear. Certainly further research along these lines is warranted.

References

Aitkin, M., & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society, Series A*, *149*, 1-43.

Bassiri, D. (1988). Large and small sample properties of maximum likelihood estimates for the hierarchical linear model. Unpublished doctoral dissertation, East Lansing, MI; Michigan State University.

Beasley, J. D. & Springer, S. G. (1977). Algorithm AS 111: The percentage points of the normal distribution. *Applied Statistics*, *26*, 118-121.

Bryk, A. S. & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, *101*, 147-158.

Bryk, A. S. & Raudenbush, S. W. (1992). *Hierarchical linear models for social and behavioral research: Applications and data analysis methods*. Beverly Hills, CA: Sage Publications.

Bryk, A. S., Raudenbush, S. W., Seltzer, M., & Congdon, R. T. (1989). *An Introduction to HLM: Computer Program and Users' Guide*. University of Chicago, Department of Education.

Coleman, J., Hoffer, T., & Kilgore, S. (1982). *High school achievement: Public, Catholic, and private schools compared*. New York: Basic Books.

Collins, L. M., Donoghue, J. R., and McGuigan, K. A. (1989, August). Assessing dependence among observations. Paper presented at the annual meeting of the American Psychological Association, New Orleans, LA.

DeLeeuw, J., & Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics*, *11*, 57-85.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, *39*, 1-18.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. D. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, *76*, 341-353.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, *73*, 43-56.

Lee, V. (1986, April). Multi-level causal models for social class and achievement. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Mason, W., Wong, G., & Entwisle, B. (1984). Contextual analysis through the multi-level linear model. In S. Leinhardt, (Ed.), *Sociological Methodology*, (pp. 72-103). San Francisco: Jossey-Bass.

Press, W. H., Flannery, B. P., Teukolsky, S. A, & Vetterling, W. T. (1989). *Numerical recipes: The art of scientific computing (FORTRAN version)*. New York: Cambridge University Press.

Raudenbush, S. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics*, *13*, 85-116.

Raudenbush, S. & Bryk, A. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics*, *10*, 75-98.

Raudenbush, S. & Bryk, A. (1986). A hierarchical model for studying school effects. *Sociology of Education*, *59*, 1-17.

Strenio, J. L. F., Weisberg, H. I., & Bryk, A. S. (1983). Empirical Bayes estimation of individual growth curve parameters and their relationship to covariates. *Biometrics*, *39*, 71-86.

Tabachnick, B. G. & Fidell, L. S. (1983). *Using multivariate statistics*. New York: Harper and Row.

Wichman, B. A. & Hill, I. D. (1982). Algorithm AS 183: An efficient and portable pseudo-random number generator. *Applied Statistics*, *31*, 188-190.

Table 1

Symbolic Specifications of Models Used
(Error Terms are Omitted from the Equations for Clarity)[3]

Correct Model

$$\beta_0 = \gamma_{00} + \gamma_{10}W_1 + \gamma_{20}W_2$$
$$\beta_1 = \gamma_{01} + \gamma_{21}W_2 + \gamma_{31}W_3$$
$$\beta_2 = \gamma_{02}$$

Errors in β_0

Correct: $\beta_0 = \gamma_{00} + \gamma_{10}W_1 + \gamma_{20}W_2$

Model *A*: $\beta_0 = \gamma_{00} + \gamma_{10}W_1 + \gamma_{20}W_2$

Model *B*: $\beta_0 = \gamma_{00} + \gamma_{10}W_1 + \gamma_{20}W_2 + \gamma_{30}W_3$

Model *C*: $\beta_0 = \gamma_{00} + \gamma_{10}W_1 + \gamma_{20}W_2 + \gamma_{30}W_3$

Errors in β_1

Correct: $\beta_1 = \gamma_{01} + \gamma_{21}W_2 + \gamma_{31}W_3$

Model *D*: $\beta_1 = \gamma_{01} + \gamma_{11}W_1 + \gamma_{21}W_2 + \gamma_{31}W_3$

Model *E*: $\beta_1 = \gamma_{01} + \gamma_{11}W_1 + \gamma_{21}W_2 + \gamma_{31}W_3$

Errors in β_2

Correct: $\beta_2 = \gamma_{02}$

Model *F*: $\beta_2 = \gamma_{02} + \gamma_{12}W_1$

Model *G*: $\beta_2 = \gamma_{02} + \gamma_{32}W_3$

Model *H*: $\beta_2 = \gamma_{02} + \gamma_{12}W_1 + \gamma_{32}W_3$

[3] Crossed out terms indicate effects which are incorrectly omitted. Underlining indicates effects which are incorrectly added.

Table 2

Characteristics of Long Runs (Iter. > 100)

	N	Mean Start log Likelihood	Mean Finish log Likelihood	Mean difference (Start-Finish)	Percent Fix-ups for Start Values
Long Runs	15	-4830.91	-4301.87	529.04	100.0
Short (NG = 300)	1425	-4396.36	-3970.06	426.30	85.8
All Other Short Runs	5760	-3871.67	-3871.39	0.28	0.2

Table 3

Mean Parameter Estimates for Correctly Specified Model

Parameter	Mean Estimate	Generating Value	Bias	MSE
γ_{00}	10.04	10.07	-0.03	.268
γ_{10}	15.28	15.27	0.01	.713
γ_{30}	5.54	5.51	0.03	.619
γ_{01}	17.52	17.52	0.00	.372
γ_{21}	-1.58	-1.56	-0.02	.545
γ_{31}	5.57	5.59	-0.02	.616
γ_{02}	3.70	3.69	0.01	.081
τ_{11}	6.61	6.60	0.01	5.388
τ_{21}	-1.44	-1.50	0.06	3.347
τ_{22}	9.51	9.40	0.11	8.645
τ_{31}	1.21	1.20	0.01	1.011
τ_{32}	-1.62	-1.60	-0.02	1.188
τ_{33}	2.49	2.50	-0.01	0.527
σ^2	9.91	9.90*	0.01	0.259

*This is an average value. The actual values depends upon the condition defining the value of the intraclass correlation. For $\rho_I = 0.3, E(\sigma^2) = 15.4$, and for $\rho_I = 0.6, E(\sigma^2) = 4.4$.

Table 4

Mean Parameter Estimates for Correct Model by Sample Composition, Identified by the Number of Groups

	Number of Groups				
	10	25	60	150	300
γ_{00}	9.88	10.05	10.11	10.09	10.07
γ_{10}	15.28	15.34	15.22	15.24	15.26
γ_{20}	5.58	5.51	5.55	5.54	5.52
γ_{01}	17.60	17.48	17.52	17.50	17.52
γ_{21}	-1.64	-1.54	-1.61	-1.57	-1.55
γ_{31}	5.68	5.44	5.57	5.59	5.56
γ_{02}	3.73	3.67	3.70	3.70	3.70
τ_{11}	6.78	6.60	6.57	6.60	6.51
τ_{21}	-1.31	-1.42	-1.40	-1.53	-1.55
τ_{22}	9.12	9.77	9.60	9.61	9.44
τ_{31}	1.16	1.21	1.22	1.23	1.24
τ_{32}	-1.63	-1.49	-1.68	-1.61	-1.61
τ_{33}	2.40	2.46	2.53	2.52	2.54
Mean d	3.04	1.54	1.00	0.65	0.52
Std.	1.54	0.51	0.32	0.21	0.14
Mean RMR	2.90	1.63	1.08	0.76	0.70
Std.	1.60	0.70	0.42	0.31	0.32

Table 5

Descriptive Statistics for Ratio:
(Mean of Estimated $se(\gamma_{mq})$) / (Observed Standard Deviation of γ_{mq} Estimates);
Computed for Correctly Specified Model

Parameter	Mean	Standard Deviation	Minimum	Maximum
γ_{00}	1.024	0.179	0.688	1.464
γ_{10}	0.905	0.235	0.406	1.328
γ_{20}	0.957	0.202	0.404	1.385
γ_{01}	1.056	0.187	0.702	1.446
γ_{21}	0.960	0.163	0.702	1.339
γ_{31}	0.963	0.208	0.498	1.299
γ_{02}	1.072	0.215	0.752	2.070

Table 6

Average Distance of Elements of γ from the Generating Values for Correct and Incorrect Model Specifications: Studied Equation

	dm_s	dc_s	difference	$t(799)$	p
β_0 is Studied Equation					
Model A	7.65	0.84	6.80	61.29	.0001
Model B	1.09	0.84	0.24	12.45	.0001
Model C	8.06	0.84	7.21	66.46	.0001
β_1 is Studied Equation					
Model D	1.24	0.92	0.31	9.33	.0001
Model E	1.31	0.92	0.38	13.81	.0001
β_2 is Studied Equation					
Model F	0.34	0.21	0.14	20.86	.0001
Model G	0.35	0.21	0.14	19.10	.0001
Model H	0.31	0.21	0.14	20.31	.0001

Table 7

Means for Salient Main Effects on Difference in Average Distance of Elements of γ from the Generating Values for Correct and Incorrect Model Specifications for Studied Equation (dm_s - dc_s): Sample Composition (Identified by the Number of Groups) and Correlation between Second-level Predictors

		Model							
		β_0 is Studied Equation			β_1 is Studied Equation		β_2 is Studied Equation		
		Model A	Model B	Model C	Model D	Model E	Model F	Model G	Model H
Number of Groups	10	7.34	0.64	9.31	0.82	0.37	0.25	0.21	0.64
	25	6.64	0.24	7.27	0.39	0.42	0.17	0.16	0.35
	60	6.66	0.14	6.64	0.16	0.34	0.12	0.09	0.22
	150	6.70	0.11	6.42	0.13	0.37	0.07	0.08	0.15
	300	6.70	0.08	6.42	0.08	0.41	0.07	0.07	0.14
Correlation between Second-level Predictors	0.3	5.05	0.24	6.24	0.23	0.34	0.14	0.15	0.28
	0.6	8.56	0.24	8.19	0.40	0.43	0.14	0.13	0.32

Table 8

Average Distance of Elements of γ from the Generating Values
for Correct and Incorrect Model Specifications: Other Equations

	dm_o	dc_o	difference	$t(799)$	p
β_0 is Studied Equation					
Model A	1.021	0.965	.056	1.76	.0788
Model B	0.945	0.965	-.020	-2.28	.0227
Model C	0.947	0.965	-.018	-2.04	.0419
β_1 is Studied Equation					
Model D	0.858	0.892	-.034	-3.21	.0014
Model E	0.992	0.892	.030	2.31	.0209
β_2 is Studied Equation					
Model F	1.302	1.320	-.018	-2.487	.0131
Model G	1.316	1.320	-.003	-0.467	.6407
Model H	1.297	1.320	-.023	-2.005	.0453

Table 9

Means for Salient Main Effects on Difference in Average Distance of Elements of γ from the Generating Values for Correct and Incorrect Model Specifications for Other Equations (dm_o - dc_o): Sample Composition (Identified by the Number of Groups)

Number of Groups	β_0 is Studied Equation			β_1 is Studied Equation		β_2 is Studied Equation		
	Model A	Model B	Model C	Model D	Model E	Model F	Model G	Model H
10	0.189	-.084	-.082	-.163	0.059	0.312	-.028	-.117
25	0.049	-.011	-.011	-.000	0.036	0.160	-.002	-.011
60	0.008	-.001	-.001	-.005	0.018	0.093	0.008	0.003
150	0.018	-.001	-.001	-.000	0.019	0.077	0.004	0.006
300	0.019	-.002	0.007	-.001	0.019	0.071	0.002	0.004

Table 10

Average bias in τ_{qq} for Correct and Incorrect Specifications

	bm_{τ}	bc_{τ}	difference	$t(799)$	p
β_0 is Studied Equation					
Model A	178.67	0.012	178.66	85.26	.0001
Model B	0.002	0.012	-0.009	-0.29	.7739
Model C	163.396	0.012	163.38	77.47	.0001
β_1 is Studied Equation					
Model D	0.142	0.108	0.034	0.80	.4213
Model E	1.791	0.108	1.684	20.78	.0001
β_2 is Studied Equation					
Model F	-0.001	-0.009	0.008	1.38	.1669
Model G	0.001	-0.009	0.010	1.38	.1692
Model H	0.014	-0.009	0.023	1.96	.0506

Table 11

Means for Salient Main Effects for ANOVA on Difference in Average bias in τ_{ii} for Correct and Incorrect Specifications: Correlation between Second-level Predictors

Correlation between Second-level Predictors	β_0 is Studied Equation			β_1 is Studied Equation		β_2 is Studied Equation		
	Model A	Model B	Model C	Model D	Model E	Model F	Model G	Model H
0.3	210.5	-0.04	199.5	0.01	2.12	0.01	0.01	0.02
0.6	146.8	0.02	127.3	0.06	1.25	0.01	0.01	0.02

Table 12

Average RMR between Elements of Estimated and Generating τ, Values for Correct and Incorrect Model Specifications: Other Equations

	$RMRm_o$	$RMRc_o$	difference	$t(799)$	p
β_0 is Studied Equation					
Model	1.551	1.398	0.154	1.12	.2635
Model B	0.399	1.398	0.001	0.17	.8865
Model C	1.407	1.398	0.009	1.78	.0758
β_1 is Studied Equation					
Model D	1.011	1.045	-0.034	-1.83	.0677
Model E	1.055	1.045	0.010	0.87	.3829
β_2 is Studied Equation					
Model F	1.772	1.796	-0.024	-1.62	.1055
Model G	1.793	1.796	-0.004	-0.41	.6823
Model H	1.761	1.796	-0.035	-1.88	.0610

Table A1

Analysis of Variance of Distance Measure d and RMR for Correctly Specified Model

		D		RMR	
Effect[1]	df	SS	η^2	SS	η^2
S	4	668.58	.600	531.83	.494
D	1	0.01	.000	1.11	.001
C	1	10.77	.010	0.23	.000
I	1	2.48	.002	0.28	.000
S	4	0.11	.000	3.84	.004
S*C	4	18.38	.017	1.58	.001
S*I	4	0.80	.001	6.94	.006
D*C	1	0.12	.000	0.00	.000
D*I	1	0.19	.000	0.51	.000
C*I	1	0.01	.000	0.34	.000
S*D*C	4	0.36	.000	3.13	.003
S*D*I	4	0.72	.001	2.84	.003
S*C*I	4	2.57	.002	0.81	.001
D*C*I	1	0.23	.000	1.66	.002
S*D*C*I	4	0.16	.000	2.88	.003
Error	760	408.17		517.88	
Total	799	1113.68		1075.85	

[1] Abbreviations used for effects: S-Sample composition, D-Heterogeneity of the within-group sizes, C-Correlation between the second-level predictors, I-value of the intraclass correlation conditional upon the model.

Appendix A2

Means and Standard Deviations of *d* for Main Effects, Correctly Specified Model

Heterogeneity	Mean	sd
Low	1.35	1.13
High	1.35	1.23

Correlation between Second-level Predictors	Mean	sd
.3	1.23	0.95
.6	1.46	1.36

Intraclass Correlation	Mean	sd
.3	1.40	1.25
.6	1.29	1.10

Sample Composition (Identified by the Number of Groups)	Mean	sd
10	1.23	1.54
25	1.54	0.51
60	1.00	0.32
150	0.65	0.20
300	0.52	0.14

Appendix A3

Means and Standard Deviations of RMR for Main Effects, Correctly Specified Model

Heterogeneity	Mean	sd
Low	1.38	0.96
High	1.45	1.33

Correlation between Second-level Predictors	Mean	sd
.3	1.40	1.04
.6	1.43	1.27

Intraclass Correlation	Mean	sd
.3	1.43	1.13
.6	1.40	1.19

Sample Composition (Identified by the Number of Groups)	Mean	sd
10	2.90	1.60
25	1.63	0.70
60	1.08	0.42
150	0.76	0.31
300	0.70	0.32